

## University of Groningen

### **A cooperative approach among methods for photometric redshifts estimation: an application to KiDS data**

Cavuoti, S.; Tortora, C.; Brescia, M.; Longo, G.; Radovich, M.; Napolitano, N. R.; Amaro, V.; Vellucci, C.; La Barbera, F.; Getman, F.

*Published in:*  
Monthly Notices of the Royal Astronomical Society

*DOI:*  
[10.1093/mnras/stw3208](https://doi.org/10.1093/mnras/stw3208)

**IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.**

*Document Version*  
Publisher's PDF, also known as Version of record

*Publication date:*  
2017

[Link to publication in University of Groningen/UMCG research database](#)

*Citation for published version (APA):*

Cavuoti, S., Tortora, C., Brescia, M., Longo, G., Radovich, M., Napolitano, N. R., Amaro, V., Vellucci, C., La Barbera, F., Getman, F., & Grado, A. (2017). A cooperative approach among methods for photometric redshifts estimation: an application to KiDS data. *Monthly Notices of the Royal Astronomical Society*, 466(2), 2039-2053. <https://doi.org/10.1093/mnras/stw3208>

#### **Copyright**

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

#### **Take-down policy**

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

# A cooperative approach among methods for photometric redshifts estimation: an application to KiDS data

S. Cavuoti,<sup>1★</sup> C. Tortora,<sup>2</sup> M. Brescia,<sup>1</sup> G. Longo,<sup>3,4</sup> M. Radovich,<sup>5</sup>  
N. R. Napolitano,<sup>1</sup> V. Amaro,<sup>3</sup> C. Vellucci,<sup>3</sup> F. La Barbera,<sup>1</sup> F. Getman<sup>1</sup>  
and A. Grado<sup>1</sup>

<sup>1</sup>*Astronomical Observatory of Capodimonte – INAF, via Moiariello 16, I-80131 Napoli, Italy*

<sup>2</sup>*Kapteyn Astronomical Institute, University of Groningen, PO Box 800, NL-9700 AV Groningen, the Netherlands*

<sup>3</sup>*Department of Physics, University Federico II, Via Cinthia 6, I-80126 Napoli, Italy*

<sup>4</sup>*California Institute of Technology – Center for Data-Driven Discovery, Pasadena, CA 91125, USA*

<sup>5</sup>*Astronomical Observatory of Padua, vicolo dell'Osservatorio 5, I-35122 Padova, Italy*

Accepted 2016 December 7. Received 2016 December 7; in original form 2016 May 2

## ABSTRACT

Photometric redshifts (photo- $z$ ) are fundamental in galaxy surveys to address different topics, from gravitational lensing and dark matter distribution to galaxy evolution. The Kilo Degree Survey (KiDS), i.e. the European Southern Observatory (ESO) public survey on the VLT Survey Telescope (VST), provides the unprecedented opportunity to exploit a large galaxy data set with an exceptional image quality and depth in the optical wavebands. Using a KiDS subset of about 25000 galaxies with measured spectroscopic redshifts, we have derived photo- $z$  using (i) three different empirical methods based on supervised machine learning; (ii) the Bayesian photometric redshift model (or BPZ); and (iii) a classical spectral energy distribution (SED) template fitting procedure (LE PHARE). We confirm that, in the regions of the photometric parameter space properly sampled by the spectroscopic templates, machine learning methods provide better redshift estimates, with a lower scatter and a smaller fraction of outliers. SED fitting techniques, however, provide useful information on the galaxy spectral type, which can be effectively used to constrain systematic errors and to better characterize potential catastrophic outliers. Such classification is then used to specialize the training of regression machine learning models, by demonstrating that a hybrid approach, involving SED fitting and machine learning in a single collaborative framework, can be effectively used to improve the accuracy of photo- $z$  estimates.

**Key words:** methods: data analysis – catalogues.

## 1 INTRODUCTION

With the advent of modern multiband digital sky surveys, photometric redshifts (photo- $z$ ) have become crucial to provide redshift estimates for the large samples of galaxies, which are required to tackle a variety of problems: weak gravitational lensing to constrain dark matter and dark energy (Kuijken et al. 2015), the identification of galaxy clusters and groups (e.g. Capozzi et al. 2009; Biviano et al. 2013; Radovich et al. 2016), the search of strong-lensing (Napolitano et al. 2016) and ultracompact galaxies (Tortora et al. 2016), as well as the study of the mass function of galaxy clusters (Albrecht et al. 2006; Peacock et al. 2006; Umetsu et al. 2012; Annunziatella et al. 2016), to quote just a few. Today, despite the initial skepticism (Baum 1962; Puschell, Owen &

Laing 1982; Koo 1985; Loh & Spillar 1986), two decades of continuous improvements of photo- $z$  techniques have led to such increase in accuracy that many ongoing and planned surveys base their core science on photo- $z$  measurements to fulfill their key scientific goals (e.g. de Jong et al. 2015; Masters et al. 2015).

The evaluation of photo- $z$  is made possible by the existence of a complex correlation among the fluxes, as measured by broad-band photometry, the spectral types of the galaxies and their distance. However, the search for the highly non-linear function that maps the photometric parameter space into the redshift one is far from trivial and can be performed in many different ways. All existing methods can be divided into two main classes: theoretical and empirical.

*Theoretical methods* use spectral energy distribution (SED) templates derived either from observed galaxy spectra or from synthetic ones. Template-based techniques are, on average, less accurate than empirical methods, but they are also free from the limitations imposed by the need of a training set. Moreover, SED fitting methods

\* E-mail: stefano.cavuoti@gmail.com

can be applied over a wide range of redshifts and intrinsic colours. They rely, however, on using a set of galaxy templates that accurately map the true distribution of galaxy SEDs (and their evolution with redshift), as well as on the assumption that the photometric calibration of the data is free from systematics. Finally, they also require a detailed understanding of how external factors, such as intergalactic and galactic extinctions, affect the final result. The templates are then shifted to any redshift in a given range and convolved with the transmission curves of the filters to create a template set for the redshift estimators (Koo 1999; Massarotti, Iovino & Buzzoni 2001a; Massarotti et al. 2001b; Csabai et al. 2003; Ilbert et al. 2006). Photometric redshifts can then be obtained by comparing observed galaxy fluxes in the  $i$ th photometric band with the library of reference fluxes, depending on (bounded by) redshift  $z$  and on a set of parameters  $T$ , that account for galaxy spectral type. For each galaxy, a  $\chi^2$  confidence test provides the values of  $z$  and  $T$  that minimize the flux residuals between observations and reference templates. A further improvement over the standard template methods was the introduction of magnitude priors defined within a Bayesian framework (BPZ; Benitez 2000), which contributes to address important information on the galaxy types and expected shape of redshift distribution.

*Empirical methods* use a Knowledge Base (hereafter KB) of objects with spectroscopically measured redshifts as a training set to obtain an empirical correlation (i.e. the mapping function) between the photometric quantities and the redshift. Empirical methods have the advantage that they do not need accurate templates, because the training set is composed of real objects, which intrinsically include effects such as the filter bandpass and flux calibration, as well as reddening. However, these methods require that the KB must provide a good coverage of the photometric space, since unreliable redshift estimates are likely to be obtained outside the colour–redshift ranges covered by the KB (Biviano et al. 2013; Brescia et al. 2013, 2015; Sanches et al. 2014; Masters et al. 2015).

Several estimators have been tested to determine the shape of the empirical mapping function, from linear or non-linear fitting (see e.g. Connolly et al. 1995) to the use of machine learning algorithms such as support vector machines (Chang & Lin 2011), artificial neural networks (McCulloch & Pitts 1943) and instance-based learning (Russell & Norvig 2003). In recent times, several attempts to combine empirical and theoretical methods as well as other approaches, based on the combination or stacking of machine learning methods, have been discussed in literature (Wolpert 1992; Carrasco Kind & Brunner 2014; Kim, Brunner & Carrasco Kind 2015; Beck et al. 2016; Fotopoulou et al. 2016; Speagle & Eisenstein 2015; Zitlau et al. 2016). Blind tests of different methods to evaluate photo- $z$  have been performed in Hogg et al. (1998) on spectroscopic data from the Keck telescope on the *Hubble Deep Field* (HDF), in Hildebrandt, Wolf & Benitez (2008) on spectroscopic data from the VIMOS VLT Deep Survey (VVDS) and the FORS Deep Field (FDF, Noll et al. 2004) on the sample of luminous red galaxies (LRGs) from the SDSS-DR6. A significant advance in comparing different methods was proposed in Hildebrandt et al. (2010), through the so-called PHAT (PHoto- $z$  Accuracy Testing) contest, which adopted the black-box approach, which is typical of proper benchmarking. They performed a homogeneous comparison of the performances, focusing the analysis on the photo- $z$  methods themselves, and setting an effective standard for the assessment of photo- $z$  accuracy.

In Cavuoti et al. (2015a), we applied an empirical method based on machine learning, i.e. the Multi Layer Perceptron with Quasi Newton Algorithm (MLPQNA, Cavuoti et al. 2012; Brescia et al. 2013, 2014, 2015), to a data set of galaxies extracted from

the Kilo Degree Survey (KiDS). The KiDS survey, thanks to the large area covered (1500 deg<sup>2</sup> at the end of the survey), the good seeing [ $\sim 0.7$  arcmin median full width at half-maximum (FWHM) in the  $r$  band] and pixel scale ( $\sim 0.2$  arcmin pixel<sup>-1</sup>), together with its depth [ $r$ -band limiting magnitude of  $\sim 25$ ;  $5\sigma$  at signal-to-noise ratio (SNR) = 5], provides large data sets of galaxies with high photometric quality in the four optical bands  $u$ ,  $g$ ,  $r$  and  $i$ , very important for accurate galaxy morphology up to  $z = 0.5$ – $0.6$ .

In this work, we apply five different photo- $z$  techniques to the same KiDS data set: (i) three empirical methods, namely: the above-mentioned MLPQNA, the Random Forest (RF; Breiman 2001) and an optimization network based on the Levenberg–Marquardt learning rule (LEMON; Nocedal & Wright 2006); (ii) the LE PHARE SED template fitting (Arnouts et al. 1999; Ilbert et al. 2006); and (iii) the Bayesian photometric redshift model (Benitez 2000). The final goal being to analyse the possibility to use these models in a cooperative way, in order to optimize the accuracy of photo- $z$  estimation.

The matching with overlapping spectroscopic surveys such as the Sloan Digital Sky Survey (SDSS; Ahn et al. 2012) and Galaxy And Mass Assembly (GAMA; Driver et al. 2011) provide a uniform and well-controlled data set to investigate (i) which method provides the most accurate photo- $z$  estimates, and (ii) whether the combination of different methods might provide useful insights into the accuracy of the final estimates.

This paper is structured as follows. In Section 2, we present the data set. The methods used to evaluate photo- $z$  are summarized in Section 3. In Section 4, we describe the experiments, and, finally, we discuss the results in Section 5. Final remarks are outlined in Section 6.

## 2 THE DATA

KiDS is an optical survey (de Jong et al. 2015), carried out with the VST-OmegaCAM camera (Kuijken 2011), dedicated mainly to studies for gravitational lensing, galaxy evolution, and searches for high- $z$  quasars and galaxy clusters. The KiDS data releases consist of tiles that are observed in the  $u$ ,  $g$ ,  $r$  and  $i$  bands. Data are processed using a distributed Oracle-based environment through the Astro-WISE (AW) optical pipeline (McFarland et al. 2013). Source extraction is performed using the algorithm KIDSCAT within the AW environment, where tile stacking, photometric calibration and astrometry are performed (see de Jong et al. 2015).

The sample of galaxies on which we performed our analysis is mostly extracted from the second data release of KiDS (KiDS-DR2; de Jong et al. 2015), which contains 148 tiles observed in all filters during the first two years of survey regular operations. We added 29 extra tiles, not included in the DR2 at the time this was released, which will be part of the forthcoming analysis. We used the multiband source catalogues, based on source detection in the  $r$ -band images. While magnitudes are measured in all filters, the star–galaxy separation as well as the positional and shape parameters is derived from the  $r$ -band data only, which typically offer the best image quality and  $r$ -band seeing  $\sim 0.65$  arcsec, thus providing the most reliable source positions and shapes. Critical areas such as saturated pixels, spikes, reflection haloes and satellite tracks are masked out, and galaxies are suitably flagged. Star–galaxy separation is based on the CLASS\_STAR (star classification) and SNR parameters provided by SExtractor (Bertin & Arnouts 1996); see also La Barbera et al. (2008) for further details about this procedure. We have retained sources with  $r$ -band SExtractor FLAGS\_r < 4, thus including objects that are very close together, very bright, with bad pixels, or blended. Further details about data reduction steps

and catalogue extraction are provided in de Jong et al. (2015) and Tortora et al. (2016).

From the original catalogue of  $\sim 22$  million sources, the star-galaxy separation leaves  $\sim 12.2$  million of galaxies. After removing those galaxies, which happen to fall in the masked regions, the final sample consisted of  $\sim 7.6$  million galaxies.

Aperture photometry in the four *ugri* bands, measured within several radii, was derived using SExtractor. In this work, we use magnitudes MAGAP\_4 and MAGAP\_6, measured within the apertures of diameters 4 and 6 arcsec, respectively. These apertures were selected to reduce the effects of seeing and to minimize the contamination from mismatched sources. To correct for residual offsets in the photometric zero-points, we used the SDSS as reference: For each KiDS tile and band, we matched bright stars with the SDSS catalogue and computed the median difference between KiDS and SDSS magnitudes (*psfMag*). For more details about data preparation and preprocessing, see de Jong et al. (2015) and Cavuoti et al. (2015a).

## 2.1 Spectroscopic base

In order to build the spectroscopic KB, we cross-matched the KiDS data with the spectroscopic samples available in the GAMA data release 2 (Driver et al. 2011; Liske et al. 2015) and SDSS-III data release 9 (Ahn et al. 2012; Bolton et al. 2012; Chen et al. 2012).

GAMA observes galaxies out to  $z = 0.5$  and  $r < 19.8$  ( $r$ -band Petrosian magnitude), by reaching a spectroscopic completeness of 98 per cent for the main survey targets. It also provides information about the quality of the redshift determination by using the probabilistically defined normalized redshift quality scale  $nQ$ . Redshifts with  $nQ > 2$  were considered the most reliable (Driver et al. 2011). For what concerns SDSS-III data, we used the low- $z$  (LOWZ) and constant-mass (CMASS) samples of the Baryon Oscillation Sky Survey (BOSS). The BOSS project obtains spectra (redshifts) for 1.5 millions of luminous galaxies up to  $z \sim 0.7$ . The LOWZ sample consists of galaxies with  $0.15 < z < 0.4$  with colours similar to LRGs, selected by applying suitable cuts on magnitudes and colours to extend the SDSS LRG sample towards fainter magnitudes/higher redshifts (see e.g. Ahn et al. 2012; Bolton et al. 2012). The CMASS sample contains three times more galaxies than the LOWZ sample, and is designed to select galaxies with  $0.4 < z < 0.8$ . The rest-frame colour distribution of the CMASS sample is significantly broader than that of the LRG one; thus, CMASS contains a nearly complete sample of massive galaxies down to  $\log M_*/M_\odot \sim 11.2$ . The faintest galaxies are at  $r = 19.5$  for LOWZ and  $i = 19.9$  for CMASS. Our matched spectroscopic sample is dominated by galaxies from GAMA (46598 versus 1618 from SDSS) at low  $z$  ( $z \lesssim 0.4$ ), while SDSS galaxies dominate the higher redshift regime (out to  $z \sim 0.7$ ), with  $r < 22$ .

## 2.2 KB definition

As a general rule, in order to avoid any possible misuse of the data, in each experiment we identified sources in the KB by adding a flag, specifying whether an object belongs to the training or test sets, respectively.

The detailed procedure adopted to obtain the two data sets used for the experiments is as follows:

(i) We excluded objects having low photometric quality (i.e. with flux error higher than one magnitude).

(ii) We removed all objects having at least one missing band (or labelled Not-a-Number or NaN), thus obtaining the clean catalogue used to create the training and test sets, in which all required photometric and spectroscopic information is complete for all objects.

(iii) We performed a randomly shuffled splitting into a training and a blind test set, by using the 60 per cent/40 per cent percentages, respectively.

(iv) We applied the following cuts on limiting magnitudes (see Cavuoti et al. 2015b for details):

- (1)  $\text{MAGAP\_4\_u} \leq 25.1$
- (2)  $\text{MAGAP\_6\_u} \leq 24.7$
- (3)  $\text{MAGAP\_4\_g} \leq 24.5$
- (4)  $\text{MAGAP\_6\_g} \leq 24.0$
- (5)  $\text{MAGAP\_4\_r} \leq 22.2$
- (6)  $\text{MAGAP\_6\_r} \leq 22.0$
- (7)  $\text{MAGAP\_4\_i} \leq 21.5$
- (8)  $\text{MAGAP\_6\_i} \leq 21.0$

(v) We selected objects with IMA\_FLAGS equal to zero in the  $g$ ,  $r$  and  $i$  bands (i.e. sources that have been flagged because located in the proximity of saturated pixels, star haloes, image border or reflections or within noisy areas, see de Jong et al. 2015). The  $u$  band is not considered in such selection since the masked regions relative to this waveband are less extended than in the other three KiDS bands.

By applying all the specified steps, the final KB consisted of 15180 training and 10067 test objects. The cuts, of course, reduce the size of the final data set for which reliable redshift estimates can be obtained; see Cavuoti et al. (2015a) for more details.

We note that, as it is well known, empirical methods can be successfully applied only within the boundaries of the input parameter space, which is properly sampled by the KB (cf. Masters et al. 2015). In other words, any bias in the KB (e.g. photometric cuts, poorly represented groups of rare and peculiar objects, etc.) is reflected also in the results. This implies that the same prescriptions applied to the KB need to be applied also to the catalogues of objects for which we derive the photo- $z$ .

## 3 THE METHODS

In this section, we shortly outline the empirical (MLPQNA, RF and LEMON) and the theoretical (LE PHARE, BPZ) methods that have been used for the comparison, which is discussed in the rest of the work.

### 3.1 The machine learning models

Among the methods that are made publicly available through the DATA Mining & Exploration Web Application RESOURCE (DAMEWARE; Brescia et al. 2014) web-based infrastructure, we picked three machine learning models: the RF (Breiman 2001), and two versions of the Multi Layer Perceptron (MLP; Rosenblatt 1961), varying in terms of backward learning methods, i.e. the Quasi Newton Algorithm (QNA; Byrd, Nocedal & Schnabel 1994) and the Levenberg-Marquardt rule (Nocedal & Wright 2006), respectively.

RF (Breiman 2001) is an ensemble learning method for classification and regression. It is a collection of simple predictors, called decision trees, where each tree is capable of producing a response to a given pattern, by subdividing the data into smaller and smaller sets based on simple decisions. The main principle behind ensemble methods is that a collection of weak learners can be joined to form



a strong learner. An RF can then be considered as a meta estimator that fits a large number of decision trees on several sub-samples of the original training set and produces an average output. Such mechanism improves the predictive accuracy, with respect to the single decision tree, and keeps overfitting under control.

LEMON (LEvenberg–Marquardt Optimization Network) is based on the modified Levenberg–Marquardt method, which makes use of the exact Hessian of the error function (and not of its linearized approximation). For networks with up to several hundreds of internal weights, this algorithm is comparable with the QNA (often faster). But its main advantage is that it does not require any stopping criterion. This method almost always converges exactly to one of the minima of a function.

The MLPQNA model, i.e. an MLP implementation with a learning rule based on the QNA, belongs to Newton’s methods specialized to find the stationary point of a function through a statistical approximation of the Hessian of the training error, obtained by a cyclic gradient calculation. MLPQNA makes use of the known L-BFGS algorithm (Limited memory – Broyden Fletcher Goldfarb Shanno; Byrd et al. 1994), originally designed for problems with a wide parameter space. The analytical details of the MLPQNA method, as well as its performances, have been extensively discussed elsewhere (Cavaletti et al. 2012; Brescia et al. 2013; Cavaletti, Brescia & Longo 2014; Cavaletti et al. 2015b).

Traditional supervised learning requires the KB to be split into training and test sets. The former is used to train the method, i.e. to infer the hidden relationship between the photometric information and the redshifts. The latter, instead, is used to evaluate – using a set of statistical estimators (see Section 3.4) – the goodness of the inferred law. To avoid biases, test and training sets are always required to have a null intersection.

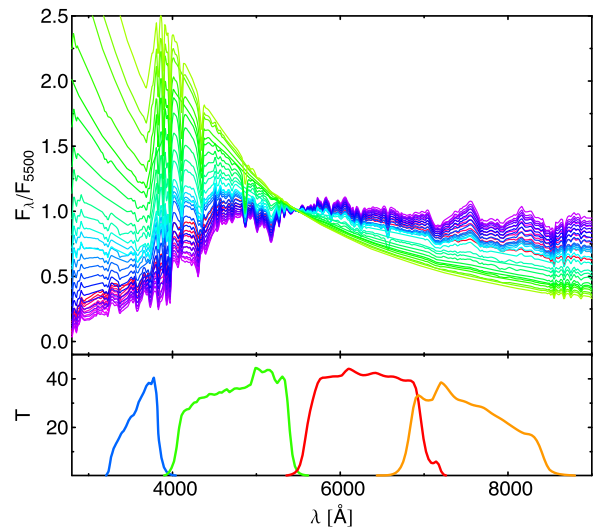
### 3.2 LE PHARE SED fitting

We use the standard SED fitting method, adopting the software LE PHARE (Arnouts et al. 1999; Ilbert et al. 2006). KiDS observed magnitudes are matched with those predicted from a set of SEDs. Each SED template is redshifted in steps of  $\delta z = 0.01$  and convolved with the four filter transmission curves. The following merit function (equation 1) is then minimized:

$$\chi^2(z, T, A) = \sum_{i=1}^{N_f} \left( \frac{F_{\text{obs}}^f - A \times F_{\text{pred}}^f(z, T)}{\sigma_{\text{obs}}^f} \right)^2, \quad (1)$$

where  $F_{\text{pred}}^f(z, T)$  is the flux predicted for a template  $T$  at redshift  $z$ .  $F_{\text{obs}}^f$  is the observed flux and  $\sigma_{\text{obs}}^f$  is the associated error derived from the observed magnitudes and errors. The index  $f$  refers to the considered filter and  $N_f = 4$  is the number of filters. The photometric redshift is determined from the minimization of  $\chi^2(z, T, A)$  varying the three free parameters  $z$ ,  $T$  and the normalization factor  $A$ . As final products of the fitting procedure, the LE PHARE code provides two main results: (i) the photometric redshift ( $z = z_{\text{phot}}$ ), and (ii) a galaxy spectral-type classification, based on the best-fitted template model  $T$ .

For the SED fitting experiments, we used the MAGAP\_6 magnitudes in the  $u$ ,  $g$ ,  $r$  and  $i$  bands (and related  $1\sigma$  uncertainties), corrected for galactic extinction using the map in Schlafly & Finkbeiner (2011). As a reference template set, we adopted the 31 SED models used for COSMOS photo- $z$  (Ilbert et al. 2009) (see Fig. 1). The basic COSMOS library is composed of nine galaxy templates from Polletta et al. (2007), which includes three SEDs of elliptical galaxies (E) and five templates of spiral galaxies (S0, Sa, Sb, Sc, Sd). These



**Figure 1.** SED templates. Flux normalized to the flux at 5000 Å versus wavelength. Templates are taken from Ilbert et al. (2006); see text for details. Redder colours are for ellipticals, blue and green for spirals and irregulars, and finally the darker green is for starburst (SB) templates. In the bottom panel, the KiDS filters are shown.

models are generated using the code GRASIL (Silva et al. 1998), providing a better joining of ultraviolet and mid-infrared than those by Coleman, Wu & Weedman (1980) used in Ilbert et al. (2006). Moreover, to reproduce very blue colours not accounted by the Polletta et al. (2007) models, 12 additional templates using Bruzual & Charlot (2003) models with starburst (SB) ages ranging from 3 to 0.03 Gyr are added. In order to improve the sampling of the redshift–colour space and therefore the accuracy of the redshift measurements, the final set of 31 spectra is obtained by linearly interpolating the original templates. We refer to it as the COSMOS library. Internal galactic extinction can be also included as a free parameter in the fitting procedure, using two different galactic extinction laws (Prevot et al. 1984; Calzetti et al. 2000), with  $E_{BV} \leq 0.5$ .

However, we followed the setup discussed in Ilbert et al. (2009), i.e. we did not apply any galactic extinction correction for models redder than the Sc templates; the galactic extinction curve provided by Prevot et al. (1984) is used for templates redder than the SB3 model, while Calzetti et al. (2000) is adopted for those bluer (including the SB3 template). Emission lines added to the templates were also implemented as discussed in Ilbert et al. (2009). Finally, LE PHARE also provides an adaptive procedure, which calculates the shifts in the photometric zero-points. The fit is first performed on the training set: The redshift is fixed to its spectroscopic value and for each waveband the code calculates average shifts that minimize the differences between observed and predicted magnitudes. This procedure is applied iteratively until convergence is reached. The offsets are then applied to the observed magnitudes of galaxies in the test sample, and the minimization of the  $\chi^2$  is performed. We tried some preliminary experiments without imposing any constraint on the fitted models, finding that about 12 per cent of the test sample would have estimated photometric redshifts larger than 1, with most of them being catastrophic outliers. For this reason, by looking at the results for the test sample, we imposed the flat prior, derived from the training data only, on absolute magnitudes. In particular, we have forced the galaxies to have absolute  $i$ -band magnitudes in the range  $(-10, -26)$ .

We also tested three different configurations: (i) the fit of SED templates with no internal galactic extinction and no emission lines; (ii) the fit of SED templates with no internal galactic extinction and no emission lines, but allowing the photometry zero-points to vary using the adaptive procedure in `LE PHARE`; (iii) the fit of SED templates, using internal galactic extinction as a free parameter, adding emission lines, and using the adaptive procedure. The best results in terms of photo-z statistical performance (see Section 3.4) were obtained with the second configuration, which is referred to hereafter as the SED fitting photo-z estimation result. The use of the spectral templates is an important part of this paper, and will be used in subsequent sections.

### 3.3 Bayesian photometric redshifts

BPZ (Benitez 2000) is a Bayesian photo-z estimation based on a template fitting method. The BPZ library is composed (Benitez et al. 2004) of four modified Coleman, Wu and Weedman types (Coleman et al. 1980), plus two Kinney, Calzetti & Bohlin (1996) SB templates. The templates include emission lines but no internal dust extinction. As recommended in the BPZ documentation, we allowed BPZ to interpolate adjacent template pairs in the colour space. If spectroscopic redshifts are available, BPZ computes the ratio of observed to model best-fitting fluxes, thus allowing us to derive a correction to the initial zero-points.

The Bayesian approach adopted in BPZ combines the likelihood that a template fits the SED of a galaxy at a given redshift with a prior defining the probability to find a galaxy of that type, as a function of magnitude and redshift. This allows us to remove those solutions that would be selected if based only on the maximum likelihood, but are in disagreement with the observed distributions. In addition to the redshift and template, BPZ also provides for each galaxy the full redshift probability distribution, and a parameter (ODDS) that provides the reliability of the solution.

### 3.4 Statistical estimators

The results were evaluated using the following set of statistical estimators for the quantity  $\Delta z = (z_{\text{spec}} - z_{\text{phot}})/(1 + z_{\text{spec}})$  on the objects in the blind test set:

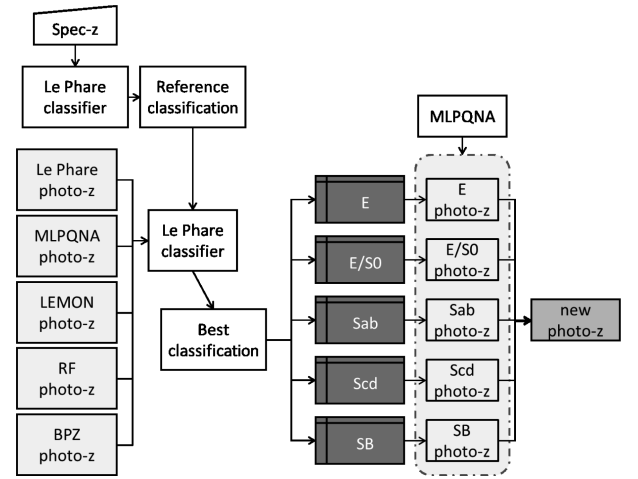
- (i) bias: defined as the mean value of the residuals  $\Delta z$ ;
- (ii)  $\sigma$ : the standard deviation of the residuals;
- (iii)  $\sigma_{68}$ : the radius of the region that includes 68 per cent of the residuals close to 0;
- (iv) NMAD: normalized median absolute deviation of the residuals, defined as  $\text{NMAD}(\Delta z) = 1.48 \times \text{Median}(|\Delta z|)$ ;
- (v) fraction of outliers with  $|\Delta z| > 0.15$ .

## 4 COMBINATION OF METHODS

The most relevant part of our work consisted of checking whether a combination of methods could be used to improve the overall results. In order to investigate such possibility, we designed a hybrid approach, which makes use of both SED fitting and Machine Learning (ML) models, organized in a workflow structured in three main stages (Fig. 2).

### 4.1 Preliminary experiments

First of all, we tested the capability of each method to deal with data affected by different systematics, e.g. photometry not corrected for



**Figure 2.** Workflow of the method implemented to combine SED fitting and ML models to improve the overall photo-z estimation quality. See text for details.

(i) galactic extinction correction; and/or (ii) the photometric zero-point offsets, as discussed in Section 2. Four experiments were performed with each model:

- (i)  $\text{EX}_{\text{clean}}$ : full KB using the clean photometry corrected by galactic extinction and offset;
- (ii)  $\text{EX}_{\text{ext}}$ : full KB using the photometry corrected by galactic extinction only (i.e. affected by an offset);
- (iii)  $\text{EX}_{\text{off}}$ : full KB using the photometry corrected by offset only (i.e. affected by galactic extinction);
- (iv)  $\text{EX}_{\text{no}}$ : full KB using the photometry not corrected by offset and galactic extinction.

SED template fitting and empirical methods are differently affected by the dereddening (i.e. the correction for galactic extinction). In the first case, in fact, reddening introduces an artificial slope in the true SED; therefore, not taking it into account would affect photometric redshift estimates. In empirical methods, instead, since it affects in the same way also the objects in the training set, it should not affect the final outcome, at least if the parameter space is properly sampled.

We need to stress that even though fitting SED templates to magnitudes not corrected for the galactic extinction is not appropriate, the inclusion/exclusion of photometric offsets and dereddening helps to quantify how the redshifts derived with different methods are affected by the presence of systematics in the photometry.

Results are summarized in Table 1 for all the experiments. In Fig. 3, we show the trends of the  $z_{\text{phot}}$  versus  $z_{\text{spec}}$  for the test objects of the  $\text{EX}_{\text{clean}}$  experiment using the five considered models, where the MLPQNA model turned out to reach the best performance among all the explored methods. Fig. 4 displays the trends of  $\Delta z$  versus  $z_{\text{spec}}$  for the same experiment and models.

### 4.2 Classification based on template fitting

The basic idea arose by analysing the photo-z estimation results on the basis of the spectral-type classification, performed by `LE PHARE` without bounding the template fitting to any redshift estimate. The statistical results summarized in Table 2 show that the machine learning models provide a better performance for all spectral types. However, ML methods perform quite differently for the different spectral types individuated by `LE PHARE`. This induced us to

**Table 1.** Blind test set statistical results for the four experiment types with the five selected methods. The outlier percentage is reported according to the rule  $|\Delta z|/(z + 1) > 0.15$ .

EXP	MLPQNA	LEMON	RF	LE PHARE	BPZ
		bias			
EX <sub>clean</sub>	0.0007	0.0006	0.0010	0.0121	0.0289
EX <sub>ext</sub>	0.0009	0.0009	0.0012	0.0183	0.0393
EX <sub>off</sub>	0.0006	0.0007	0.0010	0.0158	0.0405
EX <sub>no</sub>	0.0009	0.0010	0.0012	0.0225	0.0496
		$\sigma$			
EX <sub>clean</sub>	0.026	0.026	0.029	0.065	0.127
EX <sub>ext</sub>	0.028	0.028	0.030	0.079	0.218
EX <sub>off</sub>	0.026	0.026	0.029	0.066	0.142
EX <sub>no</sub>	0.028	0.028	0.030	0.079	0.222
		$\sigma_{68}$			
EX <sub>clean</sub>	0.018	0.018	0.021	0.041	0.039
EX <sub>ext</sub>	0.021	0.020	0.023	0.048	0.039
EX <sub>off</sub>	0.018	0.019	0.021	0.041	0.045
EX <sub>no</sub>	0.021	0.020	0.023	0.049	0.043
		NMAD			
EX <sub>clean</sub>	0.018	0.018	0.021	0.038	0.031
EX <sub>ext</sub>	0.020	0.020	0.022	0.044	0.034
EX <sub>off</sub>	0.018	0.018	0.021	0.037	0.033
EX <sub>no</sub>	0.020	0.020	0.022	0.044	0.034
		per cent outliers			
EX <sub>clean</sub>	0.31	0.30	0.40	0.89	2.18
EX <sub>ext</sub>	0.34	0.35	0.42	2.51	3.83
EX <sub>off</sub>	0.31	0.29	0.39	1.12	3.21
EX <sub>no</sub>	0.33	0.36	0.36	2.63	4.37

explore the possibility to combine the methods: Namely, the LE PHARE spectral-type classification is used to specialize ML methods and compute photo- $z$  for objects belonging to each spectral class.

Of course, the training of a specific regression model for each class can be effective only if the subdivision itself is as accurate as possible. A simple random subdivision could not enhance results. In fact, in the case of a random extraction of five subsets, the information contained in each single subset would be degraded, i.e. we would not gain any specialization but rather a reduction of patterns for each single regression network. Therefore, in this case, the best overall results would correspond to the precision achieved on the whole data set. Of course, it could happen that some subsets could improve the performance, but the overall results would be expected to remain either unchanged or get worse (Bishop 1995). Hence, we needed the best subdivision, i.e. spectral-type classification, to proceed further.

After having obtained the EX<sub>clean</sub> results, we first defined the true spectral type of each training galaxy as the best-fitting spectral type obtained with LE PHARE, constraining the redshift to its spectroscopic value. We then used LE PHARE with the five different photometric redshift estimates, thus obtaining five different spectral-type predictions for each training galaxy. The comparison of the true spectral types with the five different predictions shows that, in the absence of spectroscopic information, RF provides the most accurate spectral-type prediction.

The comparison among the different predictions is visualized (Fig. 5) by the normalized confusion matrix (Provost & Fawcett 2001). The confusion matrix is widely used to evaluate the performance of a classification: Columns represent the instances in the predicted classes (the classifier output) and rows give the expected (True) instances in the known classes. In a confusion matrix representing a two-class problem, displayed as an example in

Table 3, the quantities are TP (true positive), TN (true negative), FP (false positive) and FN (false negative). The example of a confusion matrix in Table 3 can be easily extended to the case with more than two classes: Fig. 5 shows the case of five spectral-type classes. Looking at the colour bar close to each confusion matrix panel, reddish blocks contain higher percentages of objects, while the opposite occurs for bluish blocks. The ideal condition (i.e. the perfect classification for all classes) would correspond to have red all blocks on the main diagonal of the matrix and consequently in blue the rest of the blocks. By comparing the five matrices, the RF model (panel c in Fig. 5) presents the best behaviour for all classes.

### 4.3 Redshifts for spectral-type classes

We then subdivided the KB on the base of the five spectral-type classes, thus obtaining five different subsets used to perform distinct training and blind test experiments, one for each individual class. The results for each class are depicted in Figs 6–10 and in Table 4. The figures confirm the statistical results of Table 4, where there is a clear improvement in the case of the combined approach for classes E, E/S0 and Sab, and all statistical estimators show better results than the *standard* case. A similar behaviour is visible for class Scd with the only exception of bias, while in the case of SB, all estimators are better in the combined approach, with the exception of the  $\sigma$  that remains unchanged. The resulting amount of objects for each class is obviously different from the one displayed in Table 2, which was based on a free fitting, i.e. with the model template and redshift left free to vary.

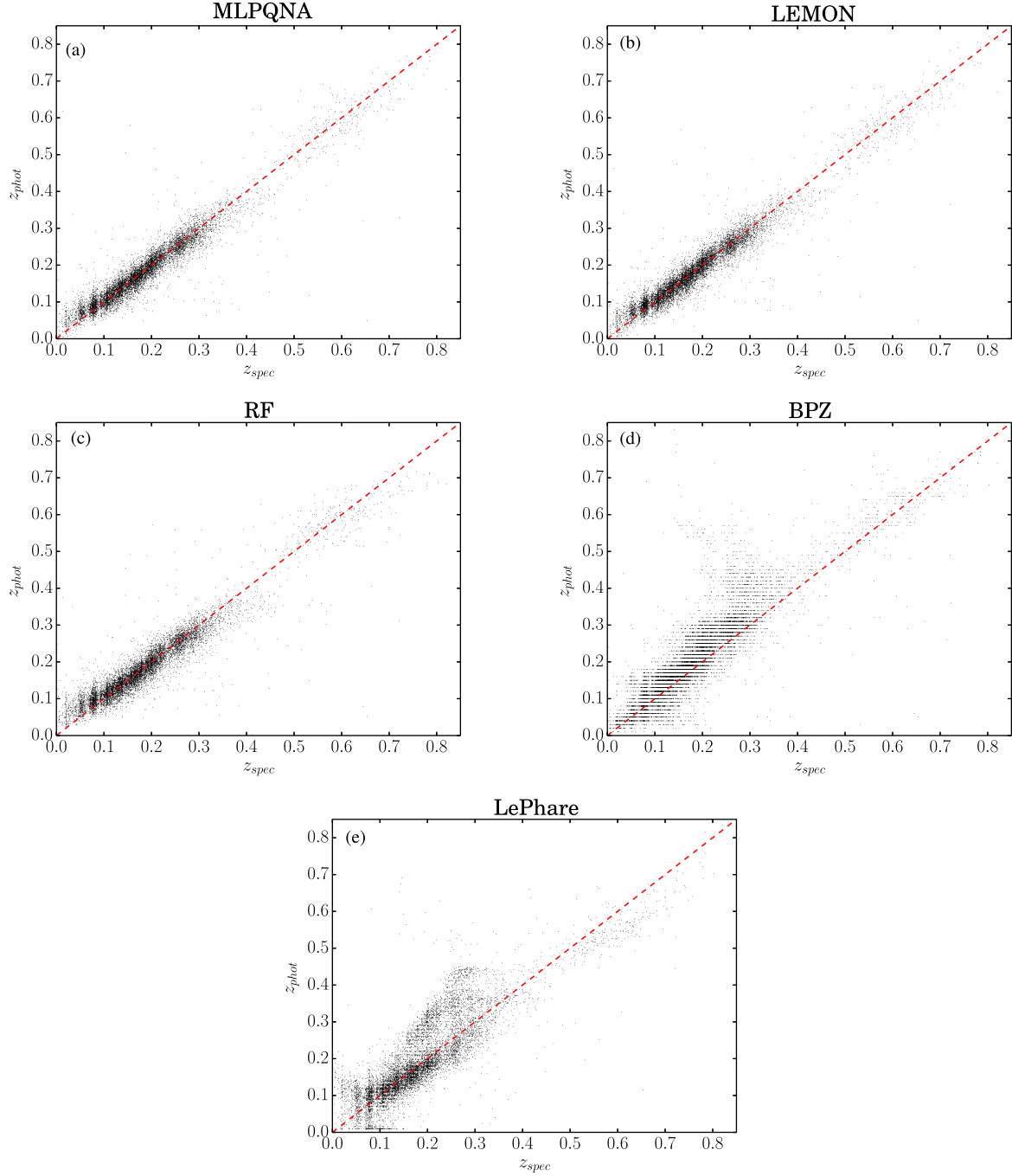
### 4.4 Recombination

The final stage of the workflow consisted of the recombination of the five subsets to produce the overall photo- $z$  estimation, which was compared with the initial EX<sub>clean</sub> experiment in terms of the usual statistical performance. By considering Table 4, the recombination statistics were calculated on the whole data sets, after having gathered together all the objects of all classes. The recombined results are reported in the last two rows of Table 4. As already emphasized for single classes, all the statistical estimators show an improvement in the combined approach case, with the exception of a slightly worse bias. Therefore, the statistics shown in Fig. 11 and in Table 4 make it apparent that the proposed combined approach induces an estimation improvement for each class as well as for the whole data set.

## 5 DISCUSSION

As discussed in Cavaletti et al. (2015a) and confirmed in Table 1, the MLPQNA outperforms SED fitting methods in all experiments. Instead, the other two empirical methods obtain results comparable with the MLPQNA. In particular, LEMON appears quite close to the MLPQNA in terms of results, a fact that could be expected by considering their similar learning laws (Shirangi & Emerick 2016). We, however, preferred MLPQNA due to its better computational efficiency.

For the EX<sub>clean</sub> experiment, we find a very small bias of  $\sim 0.0007$ , a standard deviation of  $\sim 0.026$ , a  $\sigma_{68}$  of  $\sim 0.018$ , an NMAD of  $\sim 0.018$  and a number of outliers with  $|\Delta z| > 0.15$  of only 0.31 per cent (see Table 1). In contrast, the results from SED fitting methods are less accurate, with statistical estimators and outlier fractions worse than those found using ML methods. The presence



**Figure 3.** Diagrams of  $z_{\text{spec}}$  versus  $z_{\text{phot}}$  for the data in the full redshift range available. Panels show results obtained in the case of the EX<sub>clean</sub> experiment by the various methods.

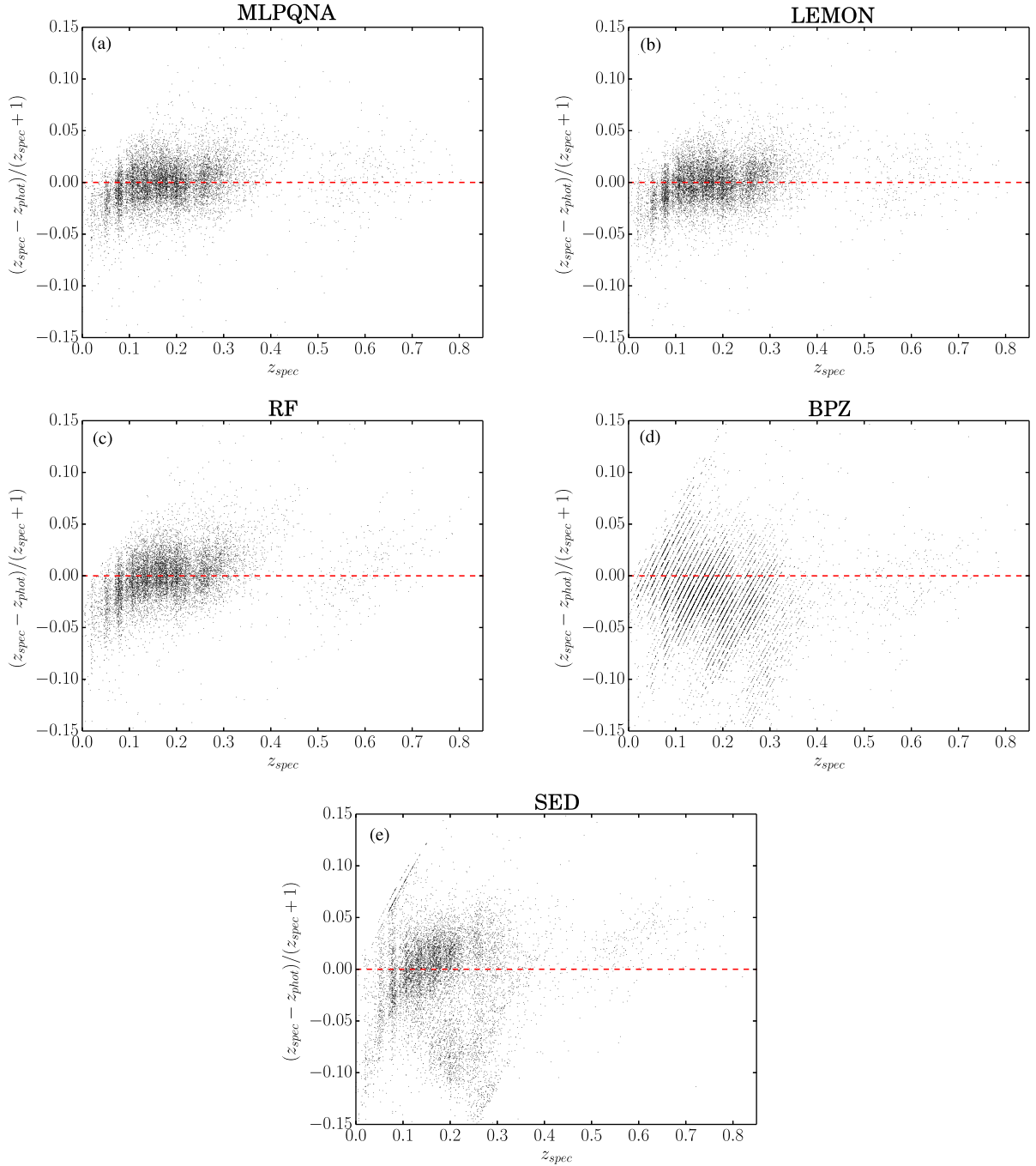
of some objects scattered around  $z_{\text{spec}} \sim 0$  confirms that there is a small residual contamination from stars misclassified as galaxies.

Furthermore, by analysing the statistics reported in Table 1, the following is evident: (i) The presence of a photometric offset (experiment EX<sub>ext</sub>) has a negligible impact on the performance of ML methods. In fact, almost all statistical estimators are the same as in the experiment with no corrections (EX<sub>no</sub>). (ii) The results of ML methods are not affected by whether the input data are dereddened or not (experiment EX<sub>off</sub>). (iii) LE PHARE is less affected by reddening than BPZ. Therefore, the main contribution to the worse performance in the experiment EX<sub>no</sub> (without offset and reddening corrections) is due to the photometric offset. In contrast, the effects

of a residual offset and reddening have a stronger impact on SED fitting methods, especially in terms of standard deviation and outliers fraction. The smaller impact on the  $\sigma_{68}$  and NMAD estimators can be justified by considering their lower dependence on the presence of catastrophic outliers, which appears as the most relevant cause of a lower performance.

The spectral-type classification provided by the SED fitting method allows us to derive also for ML models the statistical errors as a function of spectral type, thus leading to a more accurate characterization of the errors. Therefore, it is possible to assign a specific spectral-type attribute to each object and to evaluate single-class statistics. This fact, by itself, can be used to derive a better





**Figure 4.** Diagrams of  $\Delta z/(1+z)$  versus  $z_{\text{spec}}$  for the data in the full redshift range available. Panels show results obtained in the case of the EX<sub>clean</sub> experiment by the various methods.

characterization of the errors. Furthermore, as it has been shown, the combination of SED fitting and ML methods also allows to build specialized (i.e. expert) regression models for each spectral-type class, thus refining the process of redshift estimation.

During the test campaign, we also explored the possibility to increase the estimation performance by injecting the photometric redshifts calculated with LE PHARE within the parameter space used for training. But the final statistical results were slightly worse by  $\sim 1$  per cent, revealing that at least in our case such parameter does not bring enough information.

Although the spec- $z$  are in principle the most accurate information available to bound the SED fitting techniques, their use would

make impossible to produce a reliable catalogue of photometric redshifts for objects not in the KB (i.e. for objects not observed spectroscopically). Thus, it appears reasonable to identify the best solution by making use of predicted photo- $z$  to bound fitting, in order to obtain a reliable spectral-type classification for the widest set of objects. This approach, having also the capability to use arbitrary ML and SED fitting methods, makes the proposed workflow widely usable in any survey project.

By looking at Table 4, our procedure shows clearly how the MLPQNA regression method benefits from the knowledge contribution provided by the combination of the SED fitting (LE PHARE in this case) and machine learning (RF in the best case) classification

**Table 2.** Statistical results taken by considering the experiment EX<sub>clean</sub>, by distinguishing among five spectral classes of galaxies, according to the original classification performed by LE PHARE (i.e. without bounding the fitting to any kind of redshift).

	MLPQNA	LEMON	RF	LE PHARE	BPZ
class E – 2169 objects					
bias	−0.0007	−0.0004	0.0019	−0.0641	−0.0297
$\sigma$	0.022	0.022	0.024	0.045	0.041
$\sigma_{68}$	0.016	0.016	0.017	0.086	0.042
NMAD	0.015	0.015	0.016	0.036	0.027
out. (per cent)	0.18	0.23	0.28	0.60	0.65
class E/S0 – 1542 objects					
bias	0.0001	−0.0002	−0.0035	0.0124	−0.0381
$\sigma$	0.020	0.019	0.020	0.029	0.097
$\sigma_{68}$	0.014	0.014	0.016	0.0267	0.040
NMAD	0.014	0.014	0.015	0.020	0.024
out. (per cent)	0.26	0.19	0.2596	0.19	3.11
class Sab – 1339 objects					
bias	0.0007	0.0005	−0.0030	0.0073	−0.0560
$\sigma$	0.024	0.023	0.026	0.036	0.186
$\sigma_{68}$	0.019	0.020	0.023	0.030	0.050
NMAD	0.019	0.020	0.022	0.029	0.034
out. (per cent)	0.07	0.08	0.15	0.60	5.23
class Scd – 3799 objects					
bias	−0.0013	−0.0011	−0.0013	0.0022	−0.0244
$\sigma$	0.026	0.026	0.031	0.051	0.112
$\sigma_{68}$	0.020	0.019	0.023	0.028	0.036
NMAD	0.019	0.019	0.023	0.027	0.031
out. (per cent)	0.32	0.34	0.47	0.92	1.61
class SB – 1218 objects					
bias	−0.0015	−0.0012	0.0003	−0.0163	0.0005
$\sigma$	0.038	0.036	0.040	0.121	0.196
$\sigma_{68}$	0.024	0.023	0.031	0.043	0.033
NMAD	0.023	0.023	0.031	0.041	0.030
out. (per cent)	0.82	0.66	0.82	2.55	2.13

stages. In fact, this allows us to use a set of regression experts based on the MLPQNA model, specialized to predict redshifts for objects belonging to specific spectral-type classes, thus gaining in terms of a better photo-z estimation.

By analysing the results of Table 4 in more detail, the improvement in photo-z quality is significant for all classes and for all statistical estimators, as also confirmed by the comparisons in the diagrams shown in Figs 6–10. In fact, the diagrams of the residual distribution for classes E and E/S0 show a better behaviour for the combined approach in terms of distribution height and width. In the case of class Sab, the residuals of the combined approach have a more peaked distribution. Only the two classes Scd and SB show a less evident improvement, since their residual distributions appear almost comparable in both experiment types, as confirmed by their very similar values of the statistical parameters  $\sigma$  and  $\sigma_{68}$ . This leads to a more accurate photo-z prediction by considering the whole test set.

The only apparent exception is the mean (column ‘bias’ of Table 4), which suffers the effect of the alternation of positive and negative values in the *hybrid* case, which causes the algebraic sum to result slightly worse than the *standard* case (the effect occurs on the fourth decimal digit, see column ‘bias’ of the last two rows of Table 4). This is not statistically relevant because the bias is one order of magnitude smaller than  $\sigma$  and  $\sigma_{68}$  and therefore negligible.

Special attention deserves the fact that in some cases, the *hybrid* approach leads to the almost complete disappearance of catastrophic outliers. This is the case, for instance, of the E-type galaxies. The

reason is that for the elliptical galaxies the initial number of objects is lower than for the other spectral types in the KB. In the *standard* case, i.e. the standard training/test of the whole data set, such a small amount of E-type representatives are mixed together with other more populated class objects, thus causing a lower capability of the method to learn their photometric/spectroscopic correlations. Instead, in the *hybrid* case, using the proposed workflow, the possibility to learn E-type correlations through a regression expert increases the learning capabilities (see, for instance, Figs 5 and 6), thus improving the training performance and the resulting photo-z prediction accuracy.

In particular, the confusion matrices shown in Fig. 5 provide a direct visual impact and a quick comparison on the classification results. Each confusion matrix shown is referred to the results of a different spectral-type classification performed by LE PHARE, by varying the photo-z estimated through the five different regression models and used to bound the SED fitting procedure. Moreover, a confusion matrix also allows to compare classification statistics. The most important statistical estimators are (i) the *purity* or *precision*, defined as the ratio between the number of correctly classified objects of a class (the block on the main diagonal for that class) and the number of objects predicted in that class (the sum of all blocks of the column for that class); (ii) the *completeness* or *recall*, defined as the ratio between the number of correctly classified objects in that class (the block on the main diagonal for that class) and the total number of (true) objects of that class originally present in the data set (the sum of all blocks of the row for that class); and (iii) the *contamination*, automatically defined as the reciprocal value of the *purity*.

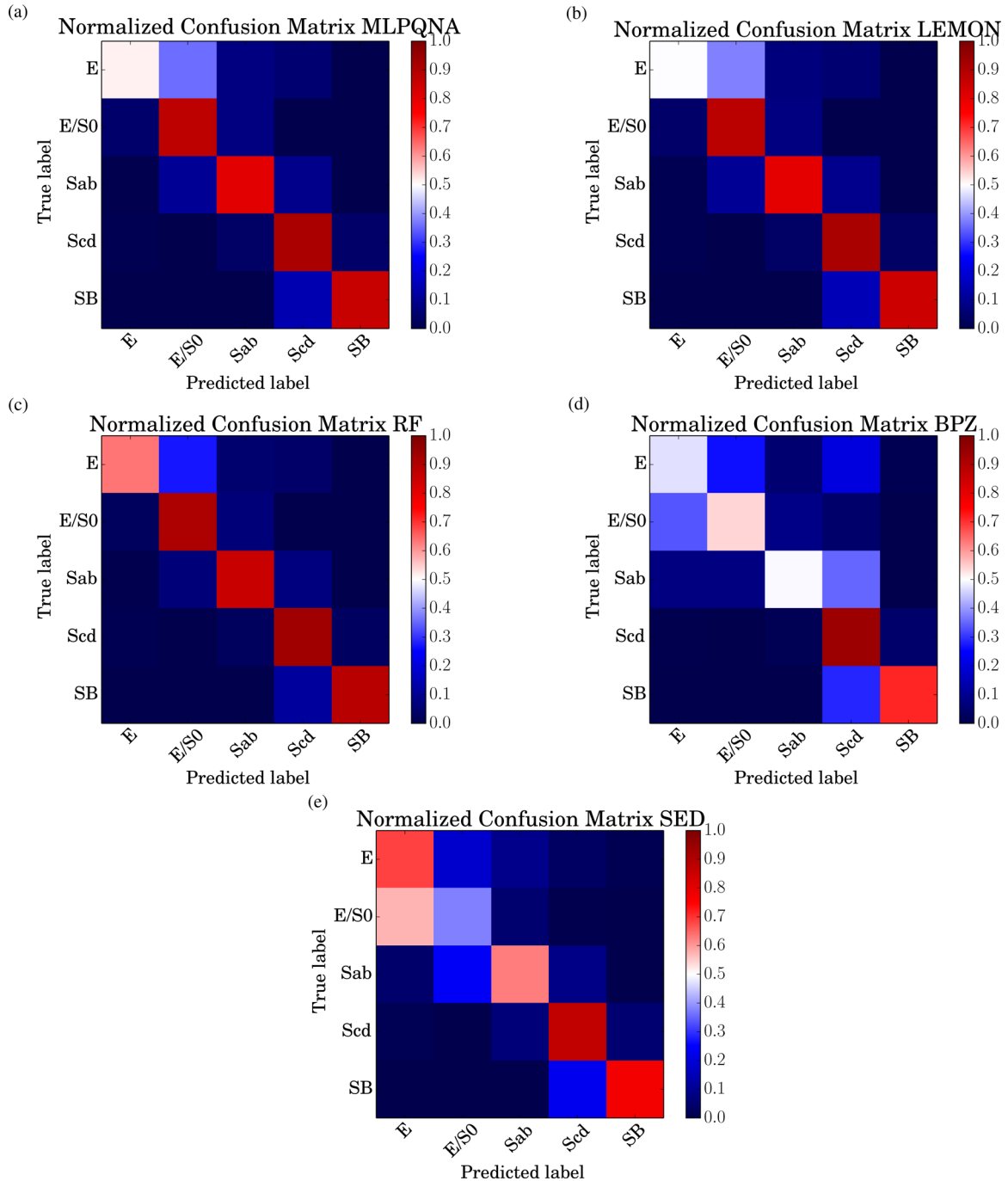
Of course, there is an obvious correspondence between the visualized colour-level confusion matrix and the *purity* and *completeness* statistics of its classes. For example, from the visual analysis of Fig. 5, it is evident that the Scd and SB spectral-type classes are well classified by all methods. This is also confirmed by their statistics, since the *purity* is, on average, around 88 per cent for Scd and 87 per cent for SB, with an averaged *completeness* of 91 per cent in the case of Scd and 82 per cent for SB.

Moreover, the confusion matrices show that the three classifications based on the machine learning models maintain a good performance in the case of the E/S0 spectral-type class, reaching on average a *purity* and a *completeness* of 89 per cent for both estimators.

In the case of the Sab class, only the RF-based classification is able to reach a sufficient degree of efficiency (78 per cent of *purity* and 85 per cent of *completeness*). In particular, for the two cases based on photo-z predicted by SED fitting models, for the Sab class, the BPZ-based results are slightly more *pure* than those based on LE PHARE (68 per cent versus 66 per cent) but much less *complete* (49 per cent versus 63 per cent).

Finally, by analysing the results on the E spectral-type class, only the RF-based case is able to maintain a sufficient compromise between *purity* (77 per cent) and *completeness* (63 per cent). The classification based on LE PHARE photo-z reaches a 69 per cent of completeness on the E class but shows an evident high level of contamination between E and E/S0, thus reducing its purity to 19 per cent. It must also be underlined that the intrinsic major difficulty to separate E objects from the E/S0 class is due to the partial co-presence of both spectral types in the class E/S0, which may partially cause wrong evaluations by the classifier.

Furthermore, the fact that latter types are less affected may be easily explained by considering that their templates are, on average, more homogeneous than for early-type objects.



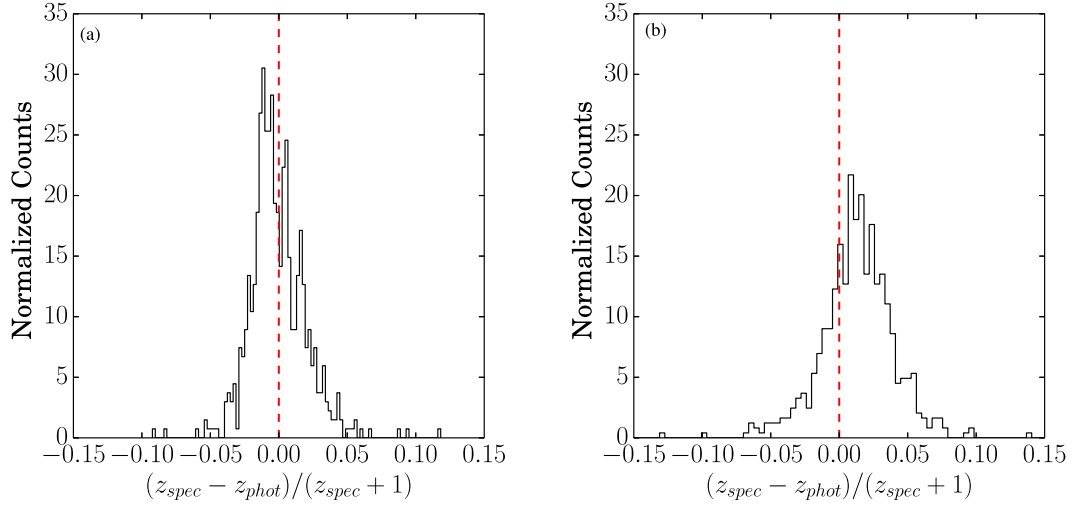
**Figure 5.** Normalized confusion matrices. The panels show LE PHARE classification results obtained by bounding the fitting with photo- $z$  derived, respectively, by (a) MLPQNA, (b) LEMON, (c) RF, (d) BPZ, and (e) LE PHARE models, based on the EX<sub>clean</sub> experiment type. Reddish blocks include higher percentages of objects, while the opposite occurs for bluish blocks. The ideal condition (perfect classification for all classes) would correspond to have red for all blocks on the main diagonal of the matrix and, consequently, in blue the rest of the blocks.

All the above considerations lead to the clear conclusion that the classification performed by the LE PHARE model and based on RF photo- $z$  achieves the best compromise between purity and completeness of all spectral-type classes. Therefore, its spectral classification has been taken as reference throughout the further steps of the workflow.

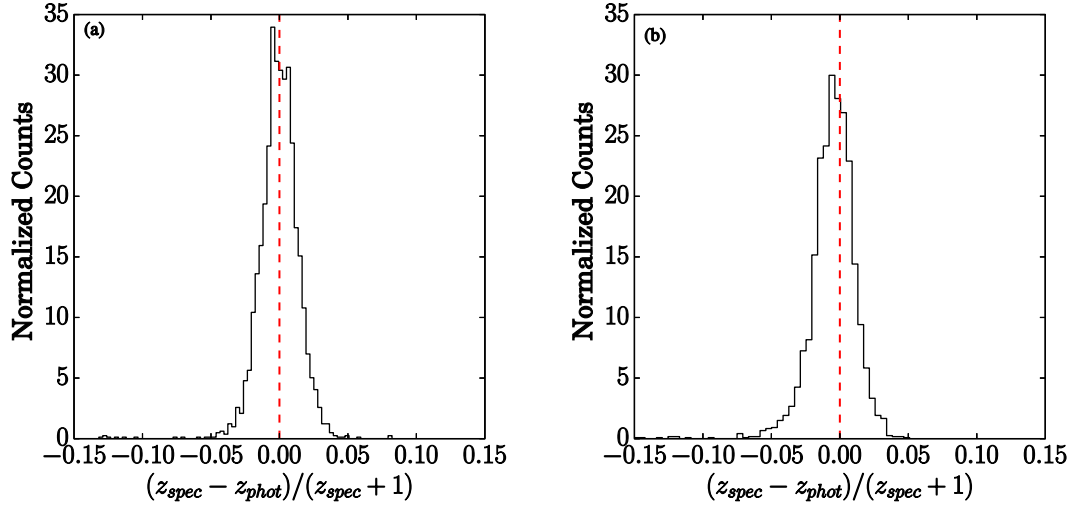
At the final stage of the proposed workflow, the photo- $z$  quality improvements obtained by the expert MLPQNA regressors on single spectral types of objects induce a reduction of  $\sigma$  from 0.026 to

**Table 3.** Structure of a confusion matrix for a two-class experiment. The interpretation of the symbols is self-explanatory. For example, TP denotes the number of objects belonging to class 1 that are correctly classified.

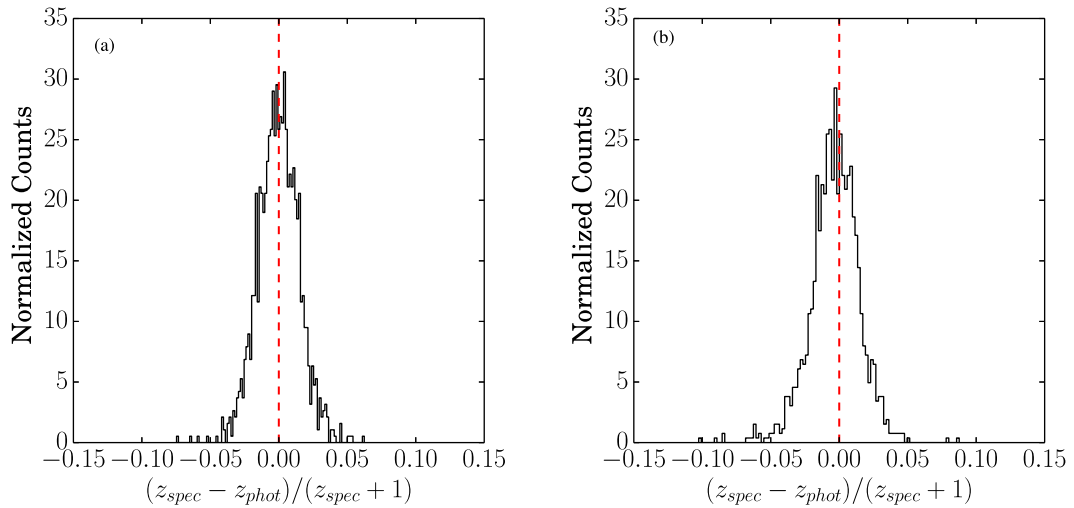
		Predicted labels	
		class 1	class 2
True Labels	class 1	TP	FN
	class 2	FP	TN



**Figure 6.** Histograms of  $\Delta z/(1+z)$  in the case of class E: The left-hand panel represents the results obtained by the *expert* MLPQNA regressor through the proposed workflow, while the right-hand panel represents the results obtained by the standard MLPQNA for the same objects.

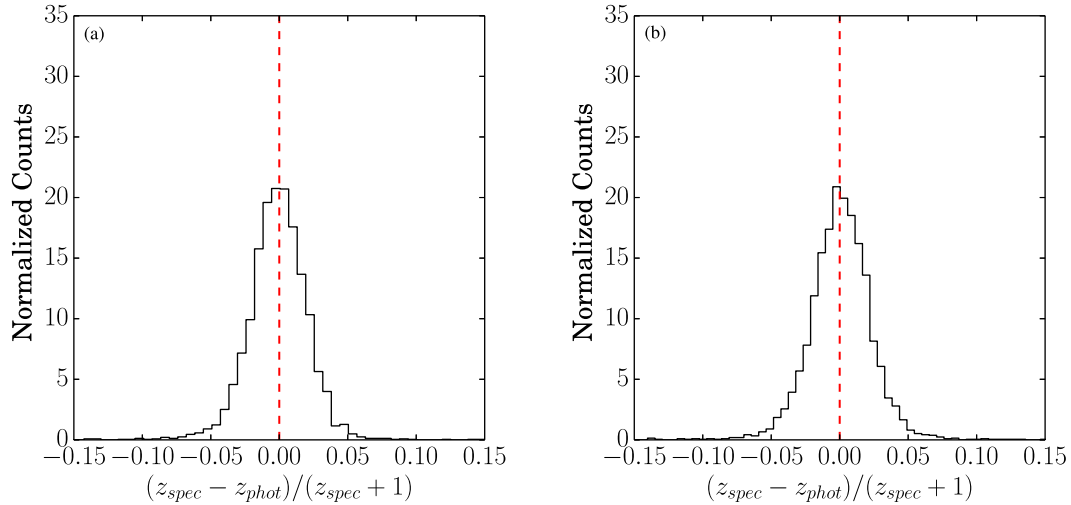


**Figure 7.** Histograms of  $\Delta z/(1+z)$  in the case of class E/S0: The left-hand panel represents the results obtained by the *expert* MLPQNA regressor through the proposed workflow, while the right-hand panel represents the result obtained by the standard MLPQNA for the same objects.

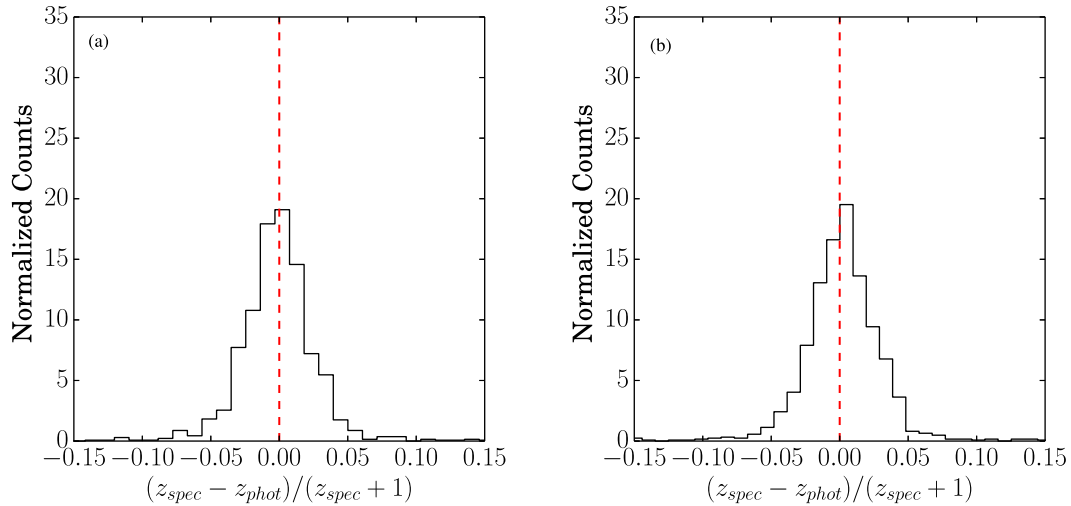


**Figure 8.** Histograms of  $\Delta z/(1+z)$  in the case of class Sab: The left-hand panel represents the results obtained by the *expert* MLPQNA regressor through the proposed workflow, while the right-hand panel represents the result obtained by the standard MLPQNA for the same objects.





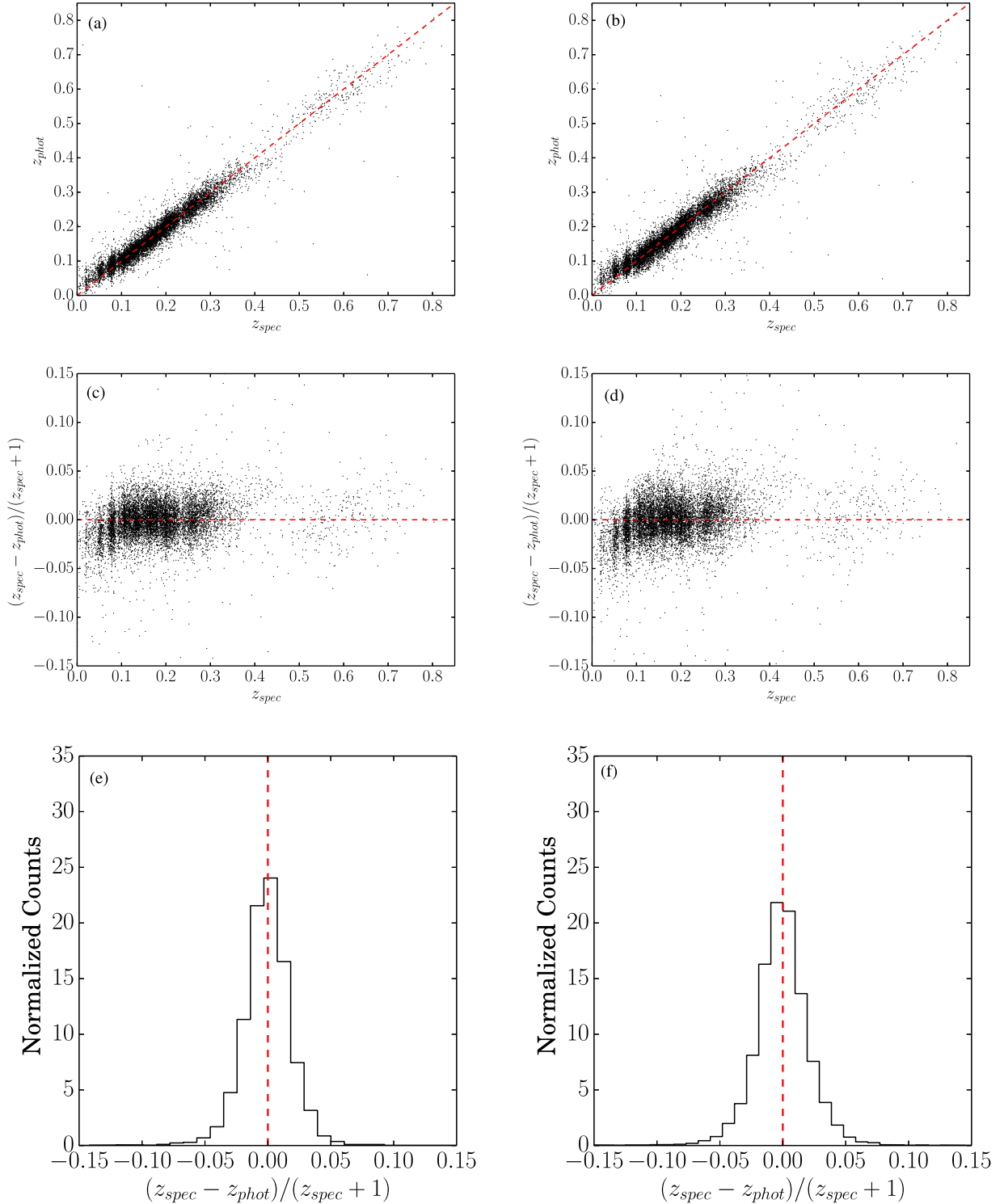
**Figure 9.** Histograms of  $\Delta z/(1+z)$  in the case of class Scd: The left-hand panel represents the results obtained by the *expert* MLPQNA regressor through the proposed workflow, while the right-hand panel represents the result obtained by the standard MLPQNA for the same objects.



**Figure 10.** Histograms of  $\Delta z/(1+z)$  in the case of class SB: The left-hand panel represents the results obtained by the *expert* MLPQNA regressor through the proposed workflow, while the right-hand panel represents the result obtained by the standard MLPQNA for the same objects.

**Table 4.** photo- $z$  estimation results, based on the MLPQNA model, for each spectral-type subset of the test set, classified by LE PHARE by bounding the fit through the photo- $z$  predicted by the RF model, which provided the best classification. The term *hybrid* refers to the results obtained by the workflow discussed here and based on the combined approach, while *standard* refers to the results obtained on the same objects but through the standard approach (i.e. EX<sub>clean</sub> experiment).

Class	Exp. type	Datasize	bias	$\sigma$	NMAD	out. (per cent)	$\sigma_{68}$
E	Hybrid	638	−0.0009	0.020	0.016	0.00	0.017
E	Standard	638	0.0130	0.029	0.022	0.31	0.028
E/S0	Hybrid	2858	−0.0005	0.016	0.012	0.10	0.012
E/S0	Standard	2858	−0.0059	0.022	0.014	0.31	0.014
Sab	Hybrid	1383	−0.0003	0.015	0.015	0.00	0.014
Sab	Standard	1383	−0.0032	0.018	0.016	0.00	0.016
Scd	Hybrid	3900	−0.0011	0.024	0.019	0.18	0.019
Scd	Standard	3900	0.0006	0.025	0.020	0.23	0.020
SB	Hybrid	1288	−0.0014	0.038	0.021	0.70	0.022
SB	Standard	1288	0.0027	0.038	0.022	0.85	0.023
ALL	Hybrid	10 067	−0.0008	0.023	0.016	0.19	0.016
ALL	Standard	10 067	−0.0007	0.026	0.018	0.31	0.018



**Figure 11.** The whole set of blind test objects: The left-hand panels represent the results obtained by grouping together all single spectral-type class outcomes of the *expert* MLPQNA regressors through the proposed *hybrid* workflow, while the right-hand column of panels represent the result obtained by the *standard* MLPQNA for the same objects. The first row shows the diagrams of  $z_{\text{spec}}$  versus  $z_{\text{phot}}$ ; the second row shows the diagrams of  $\Delta z/(1+z)$  versus  $z_{\text{spec}}$  diagrams, while the third row shows the histograms of  $\Delta z/(1+z)$ .

0.023 and of  $\sigma_{68}$  from 0.018 to 0.016 for the overall test set, besides the more relevant improvement for the E class ( $\sigma$  from 0.029 to 0.020 and of  $\sigma_{68}$  from 0.028 to 0.017). Such virtuous mechanism is mostly due to the reduction of catastrophic outliers. This signi-

ficative result, together with the generality of the workflow in terms of choice of the classification/regression methods, demonstrates the possibility to optimize the accuracy of photo- $z$  estimation through the collaborative combination of theoretical and empirical methods.

## 6 CONCLUSIONS

In this work, we propose an original workflow designed to improve the photo- $z$  estimation accuracy through a combined use of theoretical (SED fitting) and empirical (machine learning) methods.

The data sample used for the analysis was extracted from the ESO KiDS-DR2 photometric galaxy data, using a KB derived from the SDSS and GAMA spectroscopic samples. The KiDS provides wide and deep galaxy data sets with a good image quality in the optical wavebands  $u$ ,  $g$ ,  $r$  and  $i$ .

For a catalogue of about 25000 galaxies with spectroscopic redshifts, we estimated photo- $z$  using five different methods: (i) RF; (ii) MLPQNA (Multi Layer Perceptron with the Quasi Newton learning rule); (iii) LEMON (MLP with the Levenberg–Marquardt learning rule); (iv) LE PHARE SED fitting; and (v) the Bayesian model BPZ. The results obtained with the MLPQNA model on the complete KiDS-DR2 data have been discussed in Cavuoti et al. (2015a); thus, further details are provided there.

We find that, as also found in Carrasco Kind & Brunner (2014), machine learning methods provide far better redshift estimates, with a lower scatter and a smaller number of outliers when compared with the results from SED fitting. The latter, however, is able to provide very useful information on the galaxy spectral type. Such information can be effectively used to constrain the systematic errors and to better characterize potential catastrophic outliers. Furthermore, this classification can be used to specialize the training of regression machine learning models on specific types of objects. Throughout the application on KiDS data, by combining in a single collaborative framework both SED fitting and machine learning techniques, we demonstrated that the proposed workflow is capable of improving the photo- $z$  prediction accuracy.

## ACKNOWLEDGEMENTS

The authors would like to thank the anonymous referee for extremely valuable comments and suggestions. Based on data products from observations made with European Southern Observatory (ESO) telescopes at the La Silla Paranal Observatory under programme IDs 177.A-3016, 177.A-3017 and 177.A-3018, and on data products produced by Target/OmegaCEN, Istituto Nazionale di Astro Fisica (INAF)-Osservatorio Astronomico di Capodmonte Napoli (OACN), INAF-Osservatorio Astronomico di Padova (OAPD) and the KiDS production team, on behalf of the KiDS consortium. OmegaCEN and the KiDS production team acknowledge support by NOVA and NWO-M grants. Members of INAF-OAPD and INAF-OACN also acknowledge the support from the Department of Physics & Astronomy, University of Padova, and from the Department of Physics, Univ. Federico II (Naples). CT is supported through an NWO-VICI grant (project number 639.043.308). MB and SC acknowledge financial contribution from the agreement ASI/INAF I/023/12/1. MB acknowledges the PRIN-INAF 2014: ‘Glittering kaleidoscopes in the sky: the multifaceted nature and role of galaxy clusters’. GL acknowledges for partial funding from PRIN-MIUR 2011: The ‘Dark universe and the cosmic evolution of baryons: from present day surveys to Euclid’.

## REFERENCES

Ahn C. P. et al., 2012, *ApJS*, 203, 21  
 Albrecht A. et al., 2006, preprint (arXiv:astro-ph/0609591)  
 Annunziatella M. et al., 2016, *A&A*, 585, A160  
 Arnouts S. et al., 1999, *MNRAS*, 310, 540

Baum W. A., 1962, in McVittie G. C., ed., *Proc. IAU Symp. 15, Problems of Extra-Galactic Research*. Macmillan Press, New York, p. 390  
 Beck R. et al., 2016, *MNRAS*, 460, 1371  
 Benítez N., 2000, *ApJ*, 536, 571  
 Benítez N. et al., 2004, *ApJS*, 150, 1, 1  
 Bertin E., Arnouts S., 1996, *A&AS*, 117, 393  
 Bishop C. M., 1995, *Neural Networks for Pattern Recognition*. Oxford Univ. Press, Oxford  
 Biviano A. et al., 2013, *A&A*, 558, A1  
 Bolton A. S. et al., 2012, *AJ*, 144, 144  
 Breiman L., 2001, *Mach. Learn.*, 45, 25  
 Brescia M., Cavuoti S., D’Abrusco R., Mercurio A., Longo G., 2013, *ApJ*, 772, 140  
 Brescia M. et al., 2014, *PASP*, 126, 942  
 Brescia M., Cavuoti S., Longo G., De Stefano V., 2015, *A&A*, 568, A126  
 Bruzual G., Charlot S., 2003, *MNRAS*, 344, 1000  
 Byrd R. H., Nocedal J., Schnabel R. B., 1994, *Math. Program.*, 63, 129  
 Calzetti D. et al., 2000, *ApJ*, 533, 682  
 Capozzi D., De Filippis E., Paolillo M., D’Abrusco R., Longo G., 2009, *MNRAS*, 396, 900  
 Carrasco Kind M., Brunner, R. J., 2014, *MNRAS*, 442, 3380  
 Cavuoti S., Brescia M., Longo G., Mercurio A., 2012, *A&A*, 546, 13  
 Cavuoti S., Brescia M., Longo G., 2014, in Heavens A., Starck J.-L., Krone-Martins A., eds, *Proc. IAU Symp.*, 306, *Statistical Challenges in 21st Century Cosmology*. Cambridge Univ. Press, Cambridge, p. 307  
 Cavuoti S. et al., 2015, *MNRAS*, 452, 3100  
 Cavuoti S., Brescia M., De Stefano V., Longo G., 2015, *Exp. Astron.*, 39, 45  
 Chen Y.-M. et al., 2012, *MNRAS*, 421, 314  
 Chih-Chung C., Chih-Jen L., 2011, *ACM Trans. Intell. Syst. Technol.*, 2, 27  
 Coleman G. D., Wu C.-C., Weedman D. W., 1980, *ApJS*, 43, 393  
 Connolly A. J. et al., 1995, *AJ*, 110, 2655  
 Csabai I. et al., 2003, *AJ*, 125, 580  
 de Jong J. T. A. et al., 2015, *A&A*, 582, A62, 26  
 Driver S. P. et al., 2011, *MNRAS*, 413, 971  
 Fotopoulou S. et al., *MNRAS*, in press  
 Hildebrandt H., Wolf C., Benítez N., 2008, *A&A*, 480, 703  
 Hildebrandt H. et al., 2010, *A&A*, 523, 31  
 Hogg D. W., Cohen J. G., Blandford R., Pahre M. A., 1998, *ApJ*, 115, 1418  
 Ilbert O. et al., 2006, *A&A*, 457, 841  
 Ilbert O. et al., 2009, *ApJ*, 690, 1236  
 Kim E. J., Brunner R. J., Carrasco Kind M., 2015, *MNRAS*, 453, 507  
 Kinney A. L., Calzetti D., Bohlin R. C., 1996, *AJ*, 467, 38  
 Koo D. C., 1985, *AJ*, 90, 418  
 Koo D. C., 1999, in Weymann R. J., Storrie-Lombardi L. J., Sawicki M., Brunner R. J., eds, *ASP Conf. Ser. Vol. 191, Photometric Redshifts and High Redshift Galaxies*. Astron. Soc. Pac., San Francisco, p. 3  
 Kuijken K., 2011, *The Messenger*, 146, 8  
 Kuijken K. et al., 2015, *MNRAS*, 4, 3500  
 La Barbera F. et al., 2008, *PASP*, 120, 681  
 Liske J. et al., 2015, *MNRAS*, 452, 2087  
 Loh E. D., Spillar E. J., 1986, *ApJ*, 303, 154  
 Massarotti M., Iovino A., Buzzoni A., 2001a, *A&A*, 368, 74  
 Massarotti M., Iovino A., Buzzoni A., Valls-Gabaud D., 2001b, *A&A*, 380, 425  
 Masters D. et al., 2015, *ApJ*, 813, 1, 53  
 McCulloch W., Pitts W., 1943, *Bull. Math. Biophys.*, 5, 115  
 McFarland J. P. et al., 2013, *Exp. Astron.*, 35, 45  
 Napolitano N. R. et al., 2016, in Napolitano N. R., Longo G., Marconi M., Paolillo M., Iodice E., eds, *Astrophysics and Space Science Proc.*, Vol. 42, *The Universe of Digital Sky Surveys*. Springer International Publishing, Switzerland, p. 129  
 Nocedal J., Wright S. J., 2006, *Numerical Optimization*, 2nd edn. Springer, New York  
 Noll S. et al., 2004, *A&A*, 418, 885  
 Peacock J. A. et al., 2006, *ESA-ESO Working Group on Fundamental Cosmology*, Technical Report

- Polletta M. et al., 2007, *ApJ*, 663, 81
- Prevot M. L., Lequeux J., Prevot L., Maurice E., Rocca-Volmerange B., 1984, *A&A*, 132, 389
- Provost F., Fawcett T., 2001, *Machine Learning*, 42, 203
- Puschell J. J., Owen F. N., Laing R. A., 1982, *ApJ*, 257, L57
- Radovich M., Puddu E., Bellagamba F., Moscardini L., Roncarelli M., Getman F., Grado A., 2016, in Napolitano N. R., Longo G., Marconi M., Paolillo M., Iodice E., eds, *Astrophysics and Space Science Proc.*, Vol. 42, *The Universe of Digital Sky Surveys*. Springer International Publishing, Switzerland, p. 189
- Rosenblatt F., 1961, *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan Books, Washington, DC
- Russell S., Norvig P., 2003, *Artificial Intelligence: A Modern Approach*, 2nd edn. Pearson Education
- Sanches et al., 2014, *MNRAS*, 445, 1482
- Schlafly E. F., Finkbeiner D. P., 2011, *ApJ*, 737, 103
- Shirangi M. G., Emerick A. A., 2016, *J. Pet. Sci. Eng.*, 143, 258
- Silva L., Granato G. L., Bressan A., Danese L., 1998, *ApJ*, 509, 103
- Speagle J.S., Eisenstein D.J., 2015, *MNRAS*, preprint ([arXiv:1510.08073](https://arxiv.org/abs/1510.08073))
- Tortora C. et al., 2016, *MNRAS*, 457, 2845
- Umetsu K. et al., 2012, *ApJ*, 755, 1, 56
- Wolpert, D. H. 1992, *Neural Netw.*, 5, 241
- Zitlau R., Hoyle B., Paech K., Weller J., Rau M. M., Seitz S., 2016, *MNRAS*, 460, 3152

This paper has been typeset from a  $\text{\LaTeX}$  file prepared by the author.